Notes on Bates et al. (2021)

Ryan Giordano

June 8, 2023

1 Setup and problem statement

This paper takes as given a black-box algorithm $f(\cdot)$ that operates on pairs $Z_n := (X_n, Y_n)$, producing a prediction $f(X^*) := Y^*$, which we hope satisfies $\hat{Y}^* \approx Y^*$ (though we assume nothing of the form). The algorithm $f(\cdot)$ is typically constructed from a training set, which we will ignore.

We have some calibration set, $\mathcal{Z} := \{\tilde{Z}_1, \ldots, \tilde{Z}_N\}$, which we would like to use to form *interval-valued* predictions for Y^* from Y^*. That is, we will use the observations in $\mathcal{Z}$ to form a (random) set-valued function, $\mathcal{C}(X^*)$. (We could write $\mathcal{C}(X^*|\mathcal{Z})$, to emphasize the dependence on the calibration set but that would get tedious.)

I'll use $\mathcal{S}(\cdot)$ for the mapping and $\mathcal{S}$ for sets.

How to choose the mapping $\mathcal{C}(\cdot)$ to have desirable properties? We define a family of candidate sets, and a loss function describing what a "good" set looks like. Specifically, let's take

- A nested one-dimensional family of set functions, $\mathcal{C}_\lambda(\cdot)$ such that bigger λ results in bigger sets:

$$\lambda_1 < \lambda_2 \Rightarrow \mathcal{C}_{\lambda_1}(X) \subset \mathcal{C}_{\lambda_2}(X) \text{ for all } X.$$

- A loss function $\mathcal{L}(Y, \mathcal{S})$ which increases as sets get smaller:

$$\mathcal{S} \subset \mathcal{S} \Rightarrow \mathcal{L}(Y, \mathcal{S}) \geq \mathcal{L}(Y, \mathcal{S}').$$

Tension: we want small sets (small λ), but also want small $\mathcal{L}$ (big λ). This paper is all about how to choose λ to balance these desiderata with statistical guarantees.

Problem: How to use the calibration set $\mathcal{Z}$ to choose $\hat{\lambda}$ so that the loss $\mathcal{L}(Y^*, \mathcal{C}_{\hat{\lambda}}(X^*)$ is "probably" small, where "probably" accounts for randomness in both $\mathcal{Z}$ and in (X^*, Y^*)?

2 This paper's solution

Solution step one: First, the paper defines "probably small" as

$$\underset{\mathcal{Z}}{\mathbb{P}}\left(\underset{Z^*}{\mathbb{E}}\left[\mathcal{L}(Y^*, \mathcal{C}_{\hat{\lambda}}(X^*)\right] \leq \alpha\right) =: \underset{\mathcal{Z}}{\mathbb{P}}\left(\mathscr{R}(\hat{\lambda}) \leq \alpha\right) \geq 1 - \delta$$

The expectation will be called the "risk," and we'll use it enough to give it a name:

$$\mathscr{R}(\mathcal{C}_\lambda) = \mathscr{R}(\lambda) = \underset{Z^*}{\mathbb{E}}\left[\mathcal{L}(Y^*, \mathcal{C}_\lambda(X^*)\right].$$

This is not the only way to control risk, and we'll talk later about alternatives.

Now, if you knew the risk function, you would simply take

$$\lambda^* := \inf\{\lambda : \mathscr{R}(\lambda) \leq \alpha\} \text{ and } \delta = 0.$$

But we don't, so we have to estimate $\mathscr{R}(\cdot)$ using $\mathcal{Z}$. Note that we're going to both estimate the function $\lambda \mapsto \mathscr{R}(\lambda)$, and then search over our estimate to pick a λ. You might think that would require a uniform bound on the accuracy of our approximation, but it won't, due to a clever expolitation of monotonicity of the loss.

Solution step two: How to control $\mathscr{R}(\cdot)$ using $\mathcal{Z}$? The authors assume that you can form a one-sided lower confidence region for $\mathscr{R}(\lambda)$, for any λ (pointwise). That is, that you can find an upper confidence bound (UCB) $\hat{\mathscr{R}}^+(\lambda)$ such that

$$\underset{\mathcal{Z}}{\mathbb{P}}\left(\mathscr{R}(\lambda) \leq \hat{\mathscr{R}}^+(\lambda)\right) \geq 1 - \delta.$$

This UCB is all you need! There are lots of ways to construct it, using concentration inequalities, or even asymptotics (we will talk later). But once you have it, you can take

$$\hat{\lambda} := \inf\{\lambda : \hat{\mathscr{R}}^+(\lambda') < \alpha \text{ for all } \lambda' > \lambda\}.$$

Here's a proof that this works (in the case that $\mathscr{R}(\lambda)$ is continuous):

- Suppose we picked $\hat{\lambda}$ "too small:" $\hat{\lambda} < \lambda^*$, and $\mathscr{R}(\hat{\lambda}) > \mathscr{R}(\lambda^*) = \alpha$. We failed to achieve our bound — the risk at $\hat{\lambda}$ is too high.
- But we chose $\hat{\lambda}$ so that $\hat{\lambda} < \lambda^* \Rightarrow \hat{\mathscr{R}}^+(\lambda^*) < \alpha = \mathscr{R}(\lambda^*)$. In other words, the risk $\mathscr{R}(\lambda^*)$ is outside its confidence interval $(-\infty, \hat{\mathscr{R}}^+(\lambda^*))$.
- By construction $\hat{\mathscr{R}}^+(\cdot)$, this can happen with probability at most $1-\delta$. Therefore we fail to control risk with probability no more than $1-\delta$.

Solution step three: We now need only choose an UCB. Note that we can compute

$$\hat{\mathscr{R}}(\lambda) := \frac{1}{N}\sum_{n=1}^{N} \mathcal{L}(Y_n, \mathcal{C}_\lambda(X_n)).$$

For any λ, $\hat{\mathscr{R}}(\lambda)$ is an unbiased estimate of $\mathscr{R}(\lambda)$. Different concentration results of $\hat{\mathscr{R}}(\lambda)$ to its mean $\mathscr{R}(\lambda)$ can give a family of UCB.

The simplest example is Hoeffding in the case that $\mathcal{L}(\cdot) \in [0,1]$:

$$\begin{aligned}
\underset{\mathcal{Z}}{\mathbb{P}}\left(\hat{\mathscr{R}}(\lambda) - \mathscr{R}(\lambda) < -t\right) &\le \exp\left(-2Nt^2\right) = \delta \Leftrightarrow \\
\underset{\mathcal{Z}}{\mathbb{P}}\left(\mathscr{R}(\lambda) > \hat{\mathscr{R}}(\lambda) + t\right) &\le \exp\left(-2Nt^2\right) = \delta \Leftrightarrow \\
\hat{\mathscr{R}}^+(\lambda) &:= \hat{\mathscr{R}}(\lambda) + \sqrt{\ln\delta/(-2N)}.
\end{aligned}$$

This is loose, but easy to understand. For bounded losses they recommend the Waudby-Smith-Ramdas bound, which is based on a maximal inequality for martingales.

For unbounded losses, you need to assume something. They consider a Pinelis-Utev inequality ... and also simply asymptotic normal bounds.

3 Alternatives

This is not the only way to control risk. In fact, it's not much like typical conformal inference — it's more like a "statistical tolerance region" (Krishnamoorthy, 2009).

References

Bates, S. et al. (2021). “Distribution-free, risk-controlling prediction sets”. In: *Journal of the ACM (JACM)* 68.6, pp. 1–34.

Krishnamoorthy, K. (2009). *Statistical tolerance regions : theory, applications, and computation*. eng. Wiley series in probability and statistics. Hoboken, N.J: Wiley. ISBN: 9780470380260.